Module 5: Playing with RDDs

Assignment

Version 2.0

Assignment

→ Assume that you have two files. File1 structure (customer id, name, city), File2 structure (customer id, purchase_amt, date(DD-MM-YYYY format)). You have to find out (customer_name, city, total_amount, year) for the customers whose total purchase_amt is more than a threshold. You are free to create your own data and assume your own threshold

→ Find out all the records:-

» Who have the PAN CARD information, but no income information
» Who have the income information, and corresponding PANCARD entry, but without any PAN CARD number
» Records which are present only in income information, but not present in PAN CARD information

Data-set income_info & pancard_info - download link: https://edureka.wistia.com/medias/ptat8v9jj0/download?media_file_id=189237921